Strategies for Evaluating Quality Assurance Procedures

IMKE DURRE, MATTHEW J. MENNE, AND RUSSELL S. VOSE

NOAA/National Climatic Data Center, Asheville, North Carolina

(Manuscript received 22 December 2006, in final form 1 October 2007)

ABSTRACT

The evaluation strategies outlined in this paper constitute a set of tools beneficial to the development and documentation of robust automated quality assurance (QA) procedures. Traditionally, thresholds for the QA of climate data have been based on target flag rates or statistical confidence limits. However, these approaches do not necessarily quantify a procedure's effectiveness at detecting true errors in the data. Rather, as illustrated by way of an "extremes check" for daily precipitation totals, information on the performance of a QA test is best obtained through a systematic manual inspection of samples of flagged values combined with a careful analysis of geographical and seasonal patterns of flagged observations. Such an evaluation process not only helps to document the effectiveness of each individual test, but, when applied repeatedly throughout the development process, it also aids in choosing the optimal combination of QA procedures and associated thresholds. In addition, the approach described here constitutes a mechanism for reassessing system performance whenever revisions are made following initial development.

1. Introduction

Users of meteorological data may legitimately ask, "To what extent have quality assurance (QA) procedures removed significant errors from the dataset, and at what cost?" In other words, users need to know what types of errors remain in a dataset and whether the QA procedures have inadvertently removed true climate extremes. Ideally, this information would be provided via a thorough evaluation of the type-I and type-II errors (i.e., the degree to which the QA process identified good observations as erroneous and the extent to which known errors remain undetected). An assessment of the circumstances under which the two types of errors occur also benefits the user.

This paper outlines three components of such an evaluation process. The approach relies on manual inspection as a tool for 1) the selection of appropriate thresholds for individual procedures, 2) the examination of patterns in flagged values for the purpose of verifying the appropriateness of chosen thresholds, and 3) an empirical assessment of the type-I and type-II errors of a QA system. An "extremes check" for daily precipitation totals is used to illustrate the process of selecting a test threshold and verifying its appropriateness by way of a pattern analysis.

In essence, the philosophy behind the evaluation process is that it is necessary to tailor a QA system to the relevant data and to provide an empirical assessment of the system's efficiency at detecting errors in these data. The reasons for this philosophy are discussed in section 2. Some considerations determining the effectiveness of manual review as a QA evaluation technique are discussed in section 3. In sections 4–6, the three evaluation strategies are explained and illustrated. Some concluding remarks are offered in section 7.

2. Underlying philosophy

A thorough evaluation of a QA system is of importance because obvious errors sometimes remain in quality-assured datasets (e.g., see the appendix of Durre et al. 2006) and because important climatic events are occasionally identified as errors (Wolter 1997; Fiebrich and Crawford 2001; Graybeal et al. 2004a,b). For example, the QA applied to the Comprehensive Aerological Reference Dataset (CARDS) did not identify clearly erroneous pressure values, including the designation of a surface level at 70 hPa (Durre et al. 2006). Conversely, the QA system for Release 1 of the Comprehensive Ocean–Atmosphere Dataset (COADS) rejected a significant portion of the unusually warm sea

Corresponding author address: Dr. Imke Durre, NOAA/National Climatic Data Center, 151 Patton Ave., Asheville, NC 28801.
E-mail: imke.durre@noaa.gov

DOI: 10.1175/2007JAMC1706.1

surface temperatures in the central tropical Pacific Ocean during the 1982/83 El Niño event because the limits of its "trimming" check were set too tightly to accommodate both synoptic and interannual variability (Wolter 1997). These cases represent examples of type-II and type-I errors, respectively.

Typically a QA procedure is treated as a hypothesis test in which the null hypothesis is that a datum is valid. The null hypothesis is rejected when the datum (or a parameter derived from it) exceeds a specified threshold. To establish thresholds for error detection, statistical confidence limits (Collins 2001; Hubbard et al. 2005), measures of deviation from the mean (Kahl et al. 1992; Wolter 1997), or target flag rates (Graybeal et al. 2004b) have been used. The resulting thresholds often imply that an expected percentage of values will be flagged regardless of how many errors are actually in the data. If applied to error-free data, this percentage is equivalent to the type-I error rate. Otherwise, without any additional information, the type-I error rate is unknown because both data errors and valid values are likely to have been flagged.

To illustrate, consider, as in Hubbard et al. (2005), a simple test in which a value is valid only when it lays within ±3 standard deviations of the long-term mean. Assuming a normal distribution, the expectation is that 99.73% of all data values fall within these limits, yielding a flag rate of 0.27%. If no errant values exist, then all flagged values are type-I errors. In that case, the type-I error rate also is 0.27%, while the false-positive rate is 100%. This false-positive rate is of particular relevance to the data user because it means that all flagged values are valid extremes. Unfortunately, errors are usually present, in which case the false-positive rate is unknown unless flagged values are thoroughly inspected (Kunkel et al. 1998, 2005; Graybeal et al. 2004b).

Another unknown is the number of true errors that remain undetected (type-II errors). One approach to estimating the type-II error rate of a check is to introduce erroneous values into a sample dataset in order to determine whether the "seeds" are detected by the QA process (Guttman et al. 1988; Graybeal et al. 2004b; Hubbard et al. 2005). The errors introduced are either chosen to reflect known types of errors (e.g., Graybeal et al. 2004b) or are generated randomly from a uniform or normal distribution (Hubbard et al. 2005). The results provide insight into the sensitivity (power of detection) of the check versus the magnitude of error. However, when there is little knowledge of both the type and distribution of true errors, the correspondence between the miss rate for seeded errors and the miss rate for true errors is unknown. Moreover, since the values flagged are likely to be a combination of seeded errors, valid values, and true errors, error seeding is not well suited to determining a check's false-positive rate (Graybeal et al. 2004b).

From the above discussion, it follows that neither error seeding nor threshold selection based on expected error rates is sufficient for evaluating the performance of QA procedures. Rather, a thorough evaluation requires an assessment of spatial and temporal patterns in flagged observations, of biases associated with particular meteorological conditions, and of the overall type-I and type-II error rates. An evaluation should also include the determination of whether any biases or unreasonable error rates are the result of inappropriate thresholds, an inadequate representation of the spatial or temporal variability, variations in data resolution, undocumented observing practices, or systematic errors in the data.

A critical component of such an evaluation is the manual inspection of a random sample of flagged values for false positives and a randomly selected sample of all values for obvious errors that are not detected by the procedure(s). This inspection process is similar to the practice of manual validation, which is often employed in semiautomatic QA (Guttman et al. 1988; Loehrer et al. 1996; Wolter 1997; Shafer et al. 2000; Graybeal et al. 2004a). In both contexts, inspection by a human expert is used to confirm or reject the decisions made by automated procedures. However, in the approach proposed here, the purpose of the manual evaluation is to provide guidance for improvement of the system prior to deployment rather than to override a system's automated decisions during its operation. Used in this way, the manual review serves as a mechanism for empirically documenting the performance of the automatic procedures and provides the developer with considerable control over the false-positive rate of the final system.

3. Manual review as an evaluation technique

When used as a technique for assessing the performance of QA procedures during development, the manual review serves three specific purposes: the selection of test thresholds based on the frequency of false positives among the values flagged by each check, the determination of an overall false-positive rate of the QA system, and an assessment of whether significant errors remain in the quality-assured data. In each case, the goal is to estimate the frequency of errors in a certain set of data by selecting random samples of values and making a subjective decision about the validity of each value in the samples. Both the subjectivity inherent in manual review and the impact of sampling variability require that several factors be taken into account to ensure the integrity of the evaluation process.

First, the dataset used to evaluate the performance of QA procedures should reflect the full range of climatological conditions and known observing practices that are likely to be encountered once the QA procedures are deployed operationally. This ensures that the test dataset contains the full range of conditions that might lead to false positives. To also be representative of the types of errors that are likely to be present in the data to which the final system will be applied, it is desirable for the test data to originate from the same sources that will be processed operationally.

Second, the skill, accuracy, and speed with which a human validator can judge the validity of a particular value depend on the person's knowledge of meteorological data and associated reporting practices. Manual evaluation is therefore best performed by persons with experience with the range of both common and unusual conditions that are observed in a variety of climates. However, regardless of the level of expertise, it is rarely, if ever, possible to determine whether a reported value is correct. A more attainable goal is to determine whether there is sufficient evidence for an observation to be identified as an error. Criteria for forming such a judgment depend on the type of data and QA procedures being evaluated as well as on the types of independent sources of information available that might corroborate or refute the plausibility of the sample values being assessed.

A final consideration is the size of the sample used in any particular threshold selection process or system evaluation. From a statistical perspective, the evaluation of samples of data values can be viewed as a binomial experiment in which the two possible outcomes for each evaluated value are erroneous or valid. Consequently, the appropriate sample size needs to be determined on a case-by-case basis and is related to several factors. The first is the amount of time and personnel resources available for performing the evaluation. In our experience, the amount of time required for the inspection of one value averages around 5 min, but varies greatly with the number of sources of information consulted. Second, the sampling theory stipulates that the statistical uncertainty associated with the fraction of errors in a sample is directly related to the size of the sample and the true error rate in the population. The size of the population also needs to be taken into account.

4. Threshold selection technique

The type-I and type-II error rates of a QA test are directly linked to the threshold chosen for that test. For instance, a high threshold usually implies a low false-positive rate, while a lower threshold tends to detect a larger number of errors, albeit at a greater risk of overflagging. Consequently, when implementing a particular QA procedure, the threshold ideally should adhere to a target false-positive rate that reflects the desired balance between the number of errors detected and the number of false positives.

The process of selecting such a threshold is best illustrated by way of an example. Consider then a hypothetical "precipitation extremes check" that identifies erroneously large 24-h precipitation totals by comparing each observation to the overall distribution of totals at a given station and time of year. For each daily total, a ratio is calculated as

$$\text{ratio} = x/p, \tag{1}$$

where x is the daily total and p represents a specific percentile of nonzero daily values for a given site. For illustrative purposes, we use the value associated with the 95th percentile (p_{95}) for the period within a 29-day window centered on the day in question, although any reasonable percentile and window could be considered. A daily total is then flagged when the ratio exceeds a specified threshold. This threshold must be selected such that the check identifies erroneous values without flagging a significant number of real extreme events.

A logical first step in the threshold selection process is to apply the QA procedure to a representative set of data in order to identify the values that might be flagged by the check. For the precipitation extremes check, this was accomplished by setting the ratio threshold to an initial value of 1.0 and applying the check to observations from active Cooperative Observer Network (COOP) stations in the contiguous United States for the period 1961–2004. COOP data are well suited to this task because they represent a wide range of climatic conditions and because several information sources are available to enhance the manual inspection process. For example, scanned images of original COOP observer forms can be used to identify digitizing errors that might lead to erroneous precipitation extremes. Likewise, tropical cyclone track data, qualitative comparisons with surrounding stations, and reports from local newspapers and National Weather Service Forecast Offices can assist in the verification of heavy precipitation totals, thus aiding the identification of false positives.

The next step is to establish the range of parameter values within which the threshold is certain to fall. For the precipitation extremes check, this was accomplished by examining a small number of observations with ratios exceeding 1.0. The initial inspection suggested that all events with ratios greater than 30 were

TABLE 1. The QA evaluation table for the precipitation extremes checks. Results are based on the manual inspection of sample values for different ratio thresholds.

Ratio	Sample size	No. of errors in sample	False-positive rate in sample (%)	No. of values in bin	Estimated No. of false positives in bin	Cumulative No. of values	Cumulative No. of false positives	Cumulative false-positive rate (%)
≥30.0	10	10	0	114	0	114	0	0
20.0–29.9	10	10	0	56	0	170	0	0
15.0–19.9	10	10	0	55	0	225	0	0
12.0–14.9	10	8.5	15	51	8	276	8	3
10.0–11.9	10	8	20	75	15	351	23	7
9.0–9.9	10	7	30	52	16	403	39	10
8.0–8.9	10	5	50	121	61	524	100	19
7.0–7.9	10	4	60	185	111	709	211	30
6.0–6.9	10	2	80	481	385	1190	596	50

clearly erroneous and most events with ratios less than 6 were plausible because they coincided with heavy totals at neighboring stations.

The third step is to subdivide observations within the established range into reasonably sized "bins" and empirically estimate the false-positive rate within each bin. Considering the frequency distribution of the ratios in our extremes check example, the observations with ratios between 6 and 30 were grouped into the set of bins shown in Table 1. A total of 10 values were then chosen at random from each bin, the validity of each value being assessed by manually examining observations at surrounding stations as well as by consulting other sources of information when available. Each sample value that was judged to be erroneous was counted in the "number of errors in sample" column in Table 1. When a value was found to be questionable but not clearly erroneous, half an error was counted. Once a 10-value sample had been evaluated, the false-positive rate within the sample could be quantified as the percentage of sample values that were not found to be erroneous by the evaluator (e.g., 1.5 out of 10, or 15%, for the sample with ratios between 12 and 14.9).

Following the evaluation of all samples, the fourth step is to aggregate the results in order to obtain rough estimates of the false-positive rates that would be incurred for a range of ratio thresholds. First, the total number of false positives in each bin is calculated assuming, to first approximation, that each sample false-positive rate applies to all values in the corresponding bin. Second, considering the lower boundary of each bin as a potential test threshold, the "cumulative false-positive rate" for all bins above each of these thresholds is computed. That is, for a particular threshold, the cumulative false-positive rate is approximated as the ratio of the estimated total number of false positives in all bins above the threshold to the total number of values in all these bins. For example, the estimated number of false positives in the 12–14.9 bin is 15% of 51, or 8. Taking into account the bins for higher ratios, in which no false positives were found, the estimated cumulative number of false positives for a ratio threshold of 12 is also 8, while the total number of values with ratios greater than 12 is 276. Therefore, the cumulative false-positive rate is estimated to be 8 out of 276, or approximately 3%. Note that this percentage is relative to the total number of values flagged, not to the total number of data values processed.

In the final step of the threshold selection process, the threshold is chosen based on the trend in the false-positive rate as the threshold is lowered and on the types of false positives that were encountered during the evaluation process. In Table 1, the false-positive rate increases significantly for ratios below 9. Furthermore, half of all values with ratios greater than 6 are false positives, implying that the probability that a value is an error is equal to the probability that it is valid. A typical example of such a false positive is the 488 mm of rain reported at Benevides, Texas, on 11 September 1971, which corresponds to a ratio of 8.7 (i.e., the value is 8.7 times greater than the climatological 95th percentile). This total occurred in conjunction with the landfall of Hurricane Fern and is corroborated by similarly heavy totals at several nearby stations. A QA developer interested in preserving this type of extreme value could therefore set the ratio threshold to 9, leaving errors in lower ratio categories undetected by this check. If the number of true errors in the lower bins were considered to be excessive, additional checks could be developed to explicitly target those undetected errors.

5. Analysis of patterns in the flag rate

The evaluation of a QA check also requires the examination of spatial and temporal patterns of the flagged values. In theory, such patterns may be caused by concentrations of true data errors in specific regions or periods (Collins 2001; Graybeal et al. 2004a,b). On

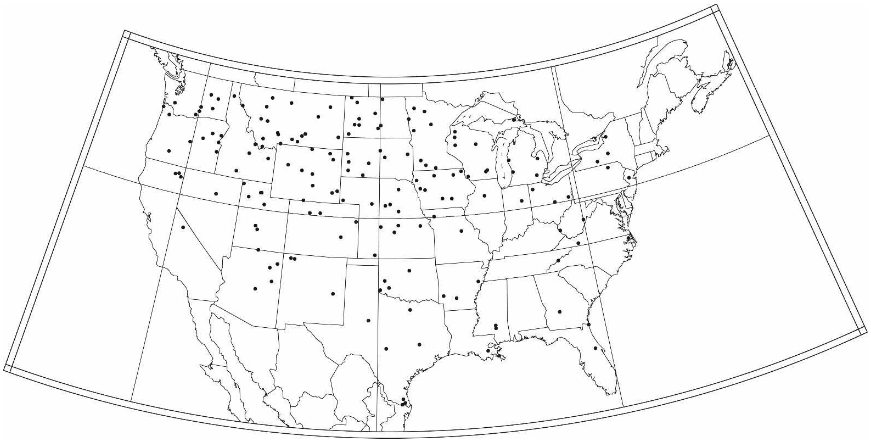

FIG. 1. Geographic distribution of stations with daily precipitation values flagged by the precipitation extremes check using a ratio threshold of 9.0 during the months November–March.

the other hand, a pattern may be indicative of procedural deficiencies such as systematic flagging of particular climatic conditions or the failure to adequately account for different observing practices (Wolter 1997; Fiebrich and Crawford 2001). Patterns may also arise as an artifact of variations in the temporal or spatial resolution of the data that limit a procedure's applicability at specific places and times (e.g., a lack of neighbors for a "spatial" consistency check).

Pattern analysis involves the generation and interpretation of summary statistics. From a QA perspective, typical examples include histograms of the overall percentage of values flagged during each calendar month and maps of stations with flagged values. An example of the latter is the spatial distribution of flags set by the precipitation extremes check during the cold season months of November–March at stations across the contiguous United States (Fig. 1). For a threshold of 9, stations with at least one flag are concentrated in the northern and interior western United States, even though the station network is less dense in this region than in other parts of the country.

An examination of supplemental information is required for determining the origin of spatial patterns of this kind. For instance, based on the examination of Cooperative Observer forms, the concentration of flags at the northern locations shown in Fig. 1 appears to result from recording a snowfall total in the water equivalent field. Thus, the pattern reflects a systematic error in the data and is not of concern. If, on the other hand, the concentration of flags had been found to be associated with overflagging of certain types of conditions, it would have been necessary to decide whether the spatial bias is acceptable or whether alterations to the QA process are necessary.

6. Analysis of QA system performance

Most QA systems consist of a suite of checks that are applied in succession (e.g., an extremes check followed by a spatial check). Once a threshold has been set for each check individually, the final step in the evaluation process should be an analysis of the overall system performance (e.g., Lorenc and Hammon 1988). The goals of such an analysis are 1) to determine whether the overall false-positive rate of the QA system matches that expected based on the thresholds chosen for each check and to estimate the corresponding type-I error rate; 2) to establish the types of errors that might remain in the processed data and whether the miss rate relative to these errors is within a reasonable bound; and 3) to assess the potential impact of any remaining errors on the likely applications of the data. All of these assessments can be made by applying the QA system to the entire dataset and manually inspecting samples of the processed data.

An effective approach to obtaining the overall false-positive and type-I error rates is to choose an appropriately sized random sample of the flagged values and then determine, via manual inspection, the number of false positives in the sample. It can also be instructive to determine the statistical uncertainty associated with the sample false-positive rate (p') thus obtained [i.e., an estimate of the degree to which p' reflects the true false-positive rate (p) within the population of all flagged values]. For example, for a sample of n values, the two-tailed hypothesis that $p = p'$ would have to be rejected at the 5% level if the number of errors in the sample fell outside the limits:

$$[n \times p - 1.96\sqrt{n \times p \times (1-p)},\; n \times p + 1.96\sqrt{n \times p \times (1-p)}]. \tag{2}$$

Such a calculation requires that p be estimated based on the results of the threshold selection process, that the sample consist of at least 20 values and constitute less than 10% of the population, and that the products $n \times p$ and $n \times (1 - p)$ both be greater than 5.

As an example, suppose that a hypothetical system consisting of five checks flags 10 000 values in a dataset of 10 million, and that the threshold for each check has been set such that the cumulative false-positive rate for bins above the threshold is 10%. It can then be estimated that the sample must contain a minimum of 5/0.1, or 50, values in order to be minimally representative of the population of all flagged values. Using Eq. (2) with $n = 50$ and $p = 0.1$, the 95% confidence limits for a sample of this size would be 5 ± 2, rounding to the nearest integer. This implies that if, for example, six of the inspected values are found to be valid during the evaluation of the 50-value sample, the hypothesis that the system's false-positive rate is 10% cannot be rejected at the 5% level. From the overall false-positive rate, the type-I error rate can then be estimated as approximately 1000 out of 10 million or 0.01%.

An analogous method for estimating the miss and type-II error rates would be to randomly choose a certain number of values from the entire dataset and manually inspect them for any remaining obvious errors. However, since clearly identifiable errors usually constitute less than 1% of the data processed (Reek et al. 1992; Kunkel et al. 1998; Graybeal et al. 2004a), the sample of values that would need to be inspected in order to obtain robust estimates of the desired metrics is likely to be excessively large given the time-consuming nature of manual inspection. Consequently, a more practical approach may be to inspect a much smaller sample of the processed data for the purpose of determining whether the miss rate in the sample is unacceptably high and results from the failure to detect a particular type of error. For example, even if only 100 randomly selected values are inspected, a finding that 1 or 2 of the sampled values are errors not identified by the automated procedures might be cause for further consideration of the overall effectiveness of the QA system. If, however, none of the values are found to be erroneous during such an inspection, the results of the inspection would be inconclusive.

To determine the impact of any remaining errors on analyses likely to be performed on the processed data, one might generate typical climatological statistics and examine them for reasonableness. If such an analysis is performed both prior to and following the application of the QA process, it can also shed light on the effectiveness of the system at removing errors that have considerable impact. For example, a dataset consisting of historical observations of daily precipitation totals for the United States is likely to be used for studies of precipitation extremes. Therefore, a suitable test would be to compute each state's highest reported precipitation total from the data and compare these extremes to corresponding record precipitation totals that have been independently verified. Prior to QA, one might expect a number of the computed extremes to far exceed the corresponding published records because of the presence of erroneous large totals. If this is still the case after QA has been applied, however, the QA system may not yet be sufficiently effective. Conversely, the finding that the extremes in the newly quality-assured data are considerably smaller than the published ones might be an indication that the QA system is overly aggressive (e.g., because the thresholds in the QA system are set too low).

7. Conclusions

The strategies outlined in this paper constitute a set of tools for evaluating automated QA procedures. These strategies, which rely heavily on manual review, are beneficial to quantifying the performance of QA checks and should be used to ensure a robust QA system. If each test is thoroughly evaluated as it is developed, the system developer has the luxury of continually adapting the QA strategy during the development process, thereby maximizing the effectiveness of the overall system once it is deployed. Without the use of these strategies, however, such control could not be exercised because neither the system's empirical false-positive rate nor its deficiencies in terms of error detection would be known.

In general, QA system development should include the following prior to deployment:

- the design of tests to detect known data problems;
- the use of manual evaluation and pattern analysis of flagged values to select test thresholds such that each check has a low false-positive rate;
- the quantification of the overall false-positive and type-I error rates for the combination of checks;
- the identification of any undetected types of errors in the quality-assured data and the assessment of the impact of such errors on likely applications of the data;
- when necessary, the development of additional checks that target undetected gross errors, followed by a reevaluation of system performance; and
- documentation of the structure and performance of the final system.

When is it necessary to develop additional checks? Although the answer may depend on the particular application for which the quality-assured data are to be used, the following general considerations may serve as guidance. First, while each individual procedure may be designed to detect only a certain kind of error (e.g., unrealistically extreme values), the entire QA system ought to identify the vast majority of egregious errors (i.e., those erroneous data points whose presence and frequency would damage the credibility of the dataset). If the miss rate is unacceptably high, the development of additional checks is likely to be more appropriate than the lowering of parameter thresholds. On the other hand, it may be necessary to abandon checks whose false-positive rates are unacceptably high for all thresholds.

Following system development, both the QA procedures and the evaluation results should be documented. At a minimum, such documentation should include a description of each check and its false-positive and flag rates as well as the type-I error rate and percentage of values flagged for the overall system. Ideally, the types of errors being detected, the types of errors that might remain, and the conditions under which valid values might be misidentified as errors would also be provided. Such documentation enables users to make informed decisions about the suitability of the data given their particular application. Furthermore, if the system is revised at a later time, the documented performance of the original system can serve as a benchmark against which the performance of the revised system can be compared.

Relative to purely statistical approaches for designing tests and choosing test thresholds, a development process that includes manual review has three distinct benefits. First, it provides the developer with considerable control over the type-I error rate of the system. As a result, the final system is likely to have a lower false-positive rate than if no evaluation were performed (e.g., Wolter 1997). Second, thanks to this robustness of the system, routine manual verification of the QA results during operation is not necessary. Third, results of the evaluation provide the end user with more detailed information about the QA system's effectiveness than can be provided with statistical measures alone. In the most general sense, these benefits make it feasible to build fully automated QA systems without diminishing the confidence in the QA decisions that are applied to the data. From an organizational perspective, this implies that if sufficient resources are committed upfront to ensure the development of a robust system, virtually no personnel resources are required once the system has been deployed, and the QA decisions for a particular set of data are reproducible.

Acknowledgments. We thank Tamara Houston, Claude Williams, and three anonymous reviewers whose thoughtful and constructive comments greatly improved this manuscript. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (Grant DE-AI02-96ER62276).

REFERENCES

Collins, W. G., 2001: The operational complex quality control of radiosonde heights and temperatures at the National Centers for Environmental Prediction. Part I: Description of the method. *J. Appl. Meteor.,* **40,** 137–151.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the integrated global radiosonde archive. *J. Climate,* **19,** 53–68.

Fiebrich, C. A., and K. C. Crawford, 2001: The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bull. Amer. Meteor. Soc.,* **82,** 2173–2187.

Graybeal, D. Y., A. T. DeGaetano, and K. L. Eggleston, 2004a: Complex quality assurance of historical hourly surface airways meteorological data. *J. Atmos. Oceanic Technol.,* **21,** 1156–1169.

——, ——, and ——, 2004b: Improved quality assurance for historical hourly temperature and humidity: Development and application to environmental analysis. *J. Appl. Meteor.,* **43,** 1722–1735.

Guttman, N. B., C. Karl, T. Reek, and V. Shuler, 1988: Measuring the performance of data validators. *Bull. Amer. Meteor. Soc.,* **69,** 1448–1452.

Hubbard, K. G., S. Goddard, W. D. Sorensen, N. Wells, and T. T. Osugi, 2005: Performance of quality assurance procedures for an Applied Climate Information System. *J. Atmos. Oceanic Technol.,* **22,** 105–112.

Kahl, J. D., M. C. Serreze, S. Shiotani, S. M. Skony, and R. C. Schnell, 1992: In-situ meteorological sounding archives for Arctic studies. *Bull. Amer. Meteor. Soc.,* **73,** 1824–1830.

Kunkel, K. E., and Coauthors, 1998: An expanded digital daily database for climatic resources applications in the Midwestern United States. *Bull. Amer. Meteor. Soc.,* **79,** 1357–1366.

——, D. R. Easterling, K. Hubbard, K. Redmond, K. Andsager, M. C. Kruk, and M. L. Spinar, 2005: Quality control of pre-1948 Cooperative Observer Network data. *J. Atmos. Oceanic Technol.,* **22,** 1691–1705.

Loehrer, S. M., T. A. Edmands, and J. A. Moore, 1996: TOGA COARE upper-air sounding data archive: Development and quality control procedures. *Bull. Amer. Meteor. Soc.,* **77,** 2651–2672.

Lorenc, A. C., and O. Hammon, 1988: Objective quality control of observations using Bayesian methods—Theory, and a practical implementation. *Quart. J. Roy. Meteor. Soc.,* **114,** 515–543.

Reek, T., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the Cooperative Observer Network. *Bull. Amer. Meteor. Soc.,* **73,** 753–762.

Shafer, M. A., C. A. Fiebrich, D. S. Arndt, S. E. Frederickson, and T. W. Hughes, 2000: Quality assurance procedures in the Oklahoma Mesonetwork. *J. Atmos. Oceanic Technol.,* **17,** 474–494.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate,* **10,** 1980–1997.